Domain-Independent Novel Event Discovery and Semi-Automatic Event Annotation

Hao Li[a], Xiang Li[a], Heng Ji[a], Yuval Marton[b]

[a]Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY 11367, USA
hengji@cs.qc.cuny.edu

[b]Center of Computational Learning Systems
Columbia University, New York, NY 10027, USA
ymarton@ccls.columbia.edu

Abstract. Information Extraction (IE) is becoming increasingly useful, but it is a costly task to discover and annotate novel events, event arguments, and event types. We exploit both monolingual texts and bilingual sentence-aligned parallel texts to cluster event triggers and discover novel event types. We then generate event argument annotations semi-automatically, framed as a sentence ranking and semantic role labeling task. Experiments on three different corpora -- ACE, OntoNotes and a collection of scientific literature -- have demonstrated that our domain-independent methods can significantly speed up the entire event discovery and annotation process while maintaining high quality.

Keywords: Information Extraction, Novel Event Discovery, Domain-Independent, Semantic Role Labeling

1 Introduction

Information Extraction (IE) techniques have been effectively applied to different domains (e.g. daily news, Wikipedia, biomedical reports, financial analysis and legal documents). A number of recent IE shared tasks (e.g. NIST Automatic Content Extraction Program (ACE 2005)) identify several common types of events. However, defining and identifying those types heavily rely on expert knowledge, and reaching an agreement among the experts or annotators requires a lot of human labor. Furthermore, annotating a high-quality event extraction corpus proves to be challenging because of highly-ambiguous and complicated event structures. The IE community has been aware of the limitation of this pre-defined event paradigm (e.g. Riloff, 1996; Yangarber et al., 2000; Grishman, 2001). Therefore a central track of IE research is the issue of *portability* – How can we automatically detect novel event types and rapidly annotate corpora for those types too, in order to alleviate the work load of human experts and annotators?

Our hypothesis is that for each domain, there are typical event types that occur frequently, and thus should arouse our interest. For example, *Transaction, Start-Organization* events are most likely to appear in *business* domain; *Justice* events are likely to appear in *law*/*criminal* domain; *Personnel* (*Start-Position* and *End-Position*) event may appear frequently in politics and business domains according to the persons involved (President Obama: politics, Bill Gates: business). The general event types defined in the existing shared tasks are far from enough to satisfy the user needs in such domains.

The first key problem is to automatically discover novel event types. In this paper we extract candidate event types based on event trigger clustering, and then rank these clusters based on their salience and novelty in the target unlabeled corpus (Section 3). After novel event types are discovered, it is beneficial to annotate arguments involved in such events. In this paper we demonstrate that such annotation can be realized in a semi-automatic way (Section 4). We take a new view of IE by considering it as a more fine-grained version of semantic role labeling (SRL). For each novel event type, we identify relevant and salient sentences involving new event triggers, and then correct errors based on uncertainty estimation, and finally map semantic roles into event argument roles based on semantic frame descriptions. We will demonstrate that our approach can make novel event discovery and annotation much more feasible (Section 5) by testing on three corpora from different domains: ACE, OntoNotes (Hovy et al., 2006) and a corpus of carbon sequestration literature (Ji et al., 2010).

2 Approach Overview

2.1 Event Definition

An event is a specific occurrence involving participants, and can be frequently described as a change of state (LDC, 2005). An event includes the following elements:

Event type: a particular event class, such as “Life/Be-Born”, “Business-Transaction”, etc.
Event trigger: the main word which most clearly expresses an event occurrence
Event arguments: the mentions that are involved in an event (participants)

For example, the sentence “*the US-led coalition troops are reportedly thrusting into the second Iraqi city of Basra*.” includes a “Movement_Transport” event that is indicated by a trigger word (“thrusting”), and a set of event arguments: the Artifact (“troops”) and the destination (“Basra”).

The challenging problem of determining the appropriate grain-size for an event (i.e. a semantic event class - sub categorization, situation types, participant roles, semantic frames…) has been debated in the IE community. In some cases the decisions should be made based on specific applications. In this paper we are not aiming to propose a new definition of events, but instead to demonstrate some general approaches which in principle could be extended to discover novel event types in any domain. We follow these two basic principles:

- Any two event triggers with the same event type should have similar distribution in text (share many neighbors in text, a.k.a. contexts), or can be translated into the same word/phrase in another language;
- The role of an event argument should be specified as much as possible, based on the event type.

2.2 Overall Pipeline

Figure 1 depicts the general procedure of our approach, which imitates a human annotator’s process by automating the following key steps:

- Automatically acquire trigger clusters and rank them according to novelty and salience in an unlabeled corpus.
- Pre-process the context sentences of novel event types by SRL, and annotate event arguments based on semantic roles semi-automatically.

The following sections will present details about these two steps respectively.

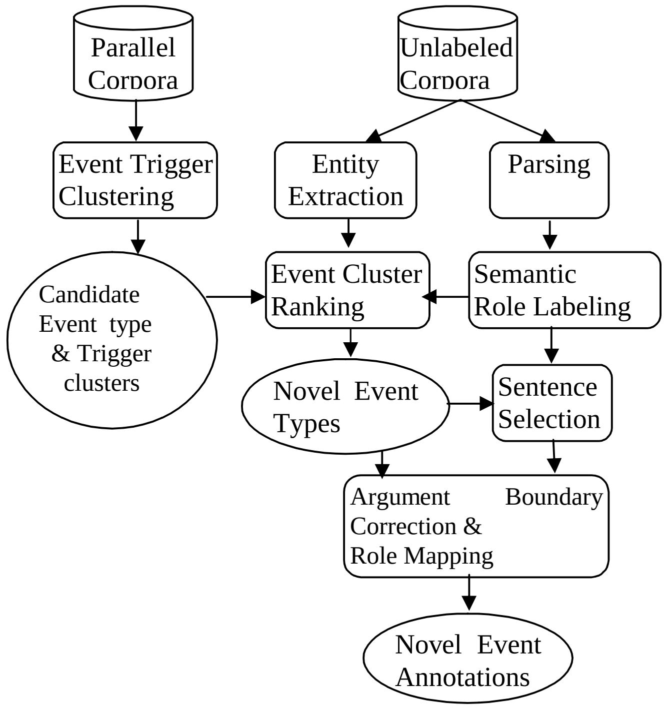

Figure 1. Novel Event Discovery and Annotation Pipeline

3 Novel Event Type Discovery

After pre-processing the text resources (Section 3.1), we continue with automatically detecting candidate event types based on trigger clustering (Sections 3.2 and 3.3), and then detecting novel event types based on cluster ranking (Section 3.4).

3.1 Pre-processing

We apply two open-domain state-of-the-art automatic trigger clustering methods to discover event trigger clusters. We consider a collection of 3065 English verbs and 4865 Chinese verbs in PropBank (Palmer et al., 2005, Xue and Palmer, 2009), together with the 1233 English triggers and 852 Chinese triggers in ACE05 training corpora as our 'pivot' event triggers. In order to minimize the impact of word alignment errors and some other noise, we conduct lemmatization based on WordNet (Fellbaum, 1998), and filter out stop-words (Fox, 1992), numbers and punctuations, time expressions and other function words that are not helpful in trigger clustering.

3.2 Monolingually derived Trigger Clustering

The first trigger clustering approach is based on distributional semantic similarity measures over a monolingual, source-language corpus (Marton et al., 2009; Marton, 2010). We constructed a monolingual English corpus of about 500 million words, consisting of all English Gigaword documents from 2004 and 2008 (LDC2009T13). With this corpus, we used essentially the earlier technique described in Marton (2010). Its outline is as follows: For each word (or word sequence) of interest *w*, collect all contexts *L w R* in which *w* appear in our monolingual corpus, and then collect paraphrase candidates: all word sequences *X* up to 6 token long appearing in the same *L X R* contexts. Then, rank the candidates *X* by their semantic similarity to *w*, as estimated by a hybrid distributional semantic similarly measure. This measure constructs a distributional profile (DP) for *w* and for each *X*, representing the DP as a vector whose dimensions correspond to the words w_i in the corpus vocabulary, and whose cell values are the log-likelihood ratio of *w* (or *X*) and w_i, as in McDonald (2000). Then, each profile is biased toward each of the word senses *s* of *w* (or sense *r* of *X*), one sense at a time, and each of the biased profiles of *w* is compared with each of the biased profiles of *X*, using cosine of these vectors. The final similarity score of *w* and *X* is the score of the highest scoring sense-biased DPs. Words senses were determined using the technique and data in Mahammad and Hirst (2006).

For each *w* we use the 50 top ranking candidates *X* above similarity score threshold 0.10, after filtering (shallowly approximated) textually entailing candidates: if all words of *w* also appear in same order in *X*, we filter *X* out. For example, for ``decided'', the candidate ``decided quickly'' is filtered out.

The semantic similarity between triggers *u* and *v* can be estimated by calculating the above similarity (vector distance) between their DPs. Then, *u* and *v* are grouped into the same cluster if the distance is shorter than a threshold. For example, an "*abandon*" cluster with similarity threshold of 0.1 includes:

{abandon, retire from, blight, quit, pull out, abducted, disuse, call off, end, withdraw, …}

3.3 Crosslingually derived Trigger Clustering

In addition to the first clustering approach, we exploit a cross-lingual clustering algorithm based on sentence-aligned bilingual parallel texts, a.k.a. bitexts (Ji, 2009) to discover additional event trigger clusters. The general idea is that if two words *w* and *u* on the bitext's source side are aligned with the same word on the target side with high confidence, then they should be grouped into the same cluster. In this paper we use Chinese-English bitexts from DARPA GALE program[1]. For each Chinese trigger, we search its aligned English words in order to construct a cluster including possible English trigger words. Then we acquire Chinese triggers from the other direction and continue the iterations. The word alignment was obtained by running Giza++ (Och and Ney, 2003). From each cluster we filter out those trigger pairs with frequency (in bitexts) less than some threshold (we have tried frequency threshold of 1, 2, 3 and 4 separately).

For example, "announce" is not an ACE-type event, but we can get its cluster as follows:

{宣布, 通告，断言，宣告，声明，预报，传达，阐明，提出，显露，显示，陈述}

→ *{announce, declare, herald, proclaim, set forth, set out, state, unveil, convey, affirm, assert}*

3.4 Event Cluster Ranking

Let *E* denote the unlabeled corpus from which we want to discover novel event types and annotate event arguments. We apply a high-performance entity extraction system (Grishman et al., 2005) and a state-of-the-art SRL system (including syntactic parsing) (Pradhan et al., 2008) to pre-process *E*.

For each candidate trigger cluster *C*, we gather it together with the entities in *E* as a query, and then use information retrieval methods to obtain *related sentences* for this query. For any word/phrase $v \in C$, if an entity *e* is identified by SRL as an argument of *v* in a sentence *s*, then *s* is the related sentence. If *v* is not tagged as a trigger for any existing event types, we consider *s* as a *novel-event related sentence*. Let $ns_k(v,e)$ be the k[th] novel-event related sentence, and $s_i(v,e)$ be the i[th] related sentence for *v* and *e*; compute the salience of C as follows.

$$salience(C,E) = \frac{\sum_{v \in C \wedge e \in E} \sum_k ns_k(v,e)}{\sum_{v \in C \wedge e \in E} \sum_i s_i(v,e)}$$

---

[1] Global Autonomous Language Exploitation

We rank all candidate trigger clusters based on their salience. The clusters with high salience values are considered as novel event types. During this procedure we also obtain a collection of candidate context sentences for each new event type. To simplify evaluation we use the most frequent trigger in each cluster as the cluster name. This is in contrast to many classical clustering algorithms, which cannot provide labels for the clusters.

4 SRL based Event Annotation

4.1 Event Extraction in SRL View

After we discover new event types, it is beneficial to annotate novel event arguments. In this paper we propose a new view of IE by considering all kinds of SRL-generated predicates with arguments (agent, object, instrument, manner, cause, property, temporal and spatial roles) as events. If we consider SRL as a simplified 'event extraction' task (considering each verb as a single event type and each argument has a general role), we may extract a temporal event chain involving "Bush" as shown in Figure 2.

ARG-TMP	*June*
Event	state
ARG0	*Bush*
ARG1	*measure*

ARG-TMP	*yesterday*
Event	declare
ARG0	*Bush*
ARG1	*meeting*

ARG-TMP	*December*
Event	summit
ARG0	*Bush*
ARG0	*Gorbachev*

Figure 2. Event Extraction in SRL view and temporal tracking

Therefore if we start from such simplified events and focus on correcting argument labeling errors, we are likely to save a large amount of time for event annotation.

4.2 Hierarchical Event Annotation

Based on the above intuitions we design the following hierarchical approach to annotate event trigger and arguments:

Step 1. Based on the salience metric defined in Section 3.4, identify a set of novel and salient event types, along with their corresponding trigger words and SRL-annotated related context sentences. Start from annotating top-ranked event types.

Step 2. Rank each sentence with novel events, which includes top-ranked clusters from step 1, according to SRL confidence score on the trigger (predicate) labeling. The lower the confidence score, the higher the sentence rank. A human annotator is then asked to correct trigger identification errors and remove irrelevant context sentences.

Step 3. Re-rank the corrected trigger-labeled sentences, obtained from step 2, according to argument difficulty. Difficulty is approximated by the number of tokens labeled with candidate roles. If this number is the same for two sentences, rank them according to sentence length.

Step 4. For each argument obtained from step 3 with SRL confidence score lower than a threshold, the human annotator is asked to manually correct the argument's boundary or role label.

Step 5. Finally, automatically map all argument roles obtained from step 4 into event argument roles based on the frame descriptions in OntoNotes. For example, the event chain in Figure 2 can be converted into Figure 3 after role mapping. For the carbon sequestration domain, a domain expert was asked to define a close set of possible roles for any specified event types. For example, a '*decrease*' event includes the following arguments: *Object, Result, Experiment, Agent, Subject, Researcher and Cause.*

5 Experimental Results

In this section we present the results of novel event discovery and the cost and quality of event annotation.

Time	*June*
Event	***announce***
Trigger	state
Announciator	*Bush*
Message	*measure*

Time	*yesterday*
Event	***announce***
Trigger	declare
Announciator	*Bush*
Message	*meeting*

Time	*December*
Event	***meeting***
Trigger	summit
Person	*Bush*
Person	*Gorbachev*

Figure 3. Event Extraction after trigger clustering and role mapping from SRL

5.1 Data and Scoring Metric

In order to test how robust our approach is, we evaluated our methods on three different data sets as shown in Table 1.

Table 1. Target Corpus Profile

Corpus	#docs	SRL	Purpose
ACE	106	Automatic	To evaluate trigger clustering and event argument annotation quality
OntoNotes	516	Ground-truth	To measure the impact of SRL errors
Carbon Sequestration Literature	150	Automatic	To evaluate portability of our methods to a non-news domain

In addition, we used English Gigaword 2004 and 2008 corpora for the monolingually derived trigger clustering method (Section 3.2), and a Chinese-English parallel corpus including 200,000 sentence pairs (part of Global Autonomous Language Exploitation Y3 Machine Translation training corpora) for the crosslingual clustering method (Section 3.3).

We evaluate event argument annotation with respect to both annotation cost (time) and quality on two data sets: ACE05 event training corpus, and the carbon sequestration literature corpus with ground-truth event annotation. We define the following standards to determine the *correctness* of an event:

- *A trigger is correctly labeled* if its event type and offsets match a reference trigger.
- *An argument is correctly labeled* if its event type, offsets, and role match any of the reference argument mentions.

5.2 Trigger Clustering Performance

We obtained 2435 clusters (average size of each cluster is 13 words) from monolingually derived trigger clustering; and 3602 clusters (average size of each cluster is 14 words) from crosslingually derived trigger clustering. In order to evaluate clustering quality, we automatically aligned our clusters with the 33 ACE05 trigger clusters by computing maximal overlaps on cluster members.

Table 2 compares the overall Precision, Recall and F-measure for these two methods when they use optimized filtering thresholds. Table 3 shows purity of the two methods. Comparing with the results of this two clustering methods, we can see that under the optimized thresholds, the crosslingual clustering method achieved significantly higher precision than the monolingual clustering method (23.65% absolute gain) with some small loss in recall (4.06%). It also achieved higher (22.21%) purity.

Table 2. Trigger Clustering Performance (%)

Method	Precision	Recall	F-Measure
Monolingual clustering	45.59	40.18	33.67
Crosslingual clustering	69.21	36.12	42.05

Table 3. Trigger Clustering Purities (%)

Method	Purity
Monolingual clustering	39.26
Crosslingual clustering	61.47

We found that although most ACE event types rank very high, a lot of other important event types are missing in the ACE paradigm. This observation exactly matches the motivation for the needs of automatic discovery of novel event types.

The top-ranked clusters for three different target corpora are as follows: "*Personnel-election*" for ACE corpus, "*Support*" for OntoNotes and "*Decrease*" for the carbon sequestration literature corpus.

It is interesting to see that the ranking results are very different between the news domain and the non-news domain (the carbon sequestration corpus). There are a lot of "*decrease*" events in the carbon sequestration domain because many experiments in this domain involve "*decrease*" as goals, such as "decreasing the emissions of CO2 gas", "decrease the concentration of CO2 in the atmosphere", "decreasing the long-term sequestration of fresh plant-C inputs into the soils" and "a decrease in grazing intensity".

5.3 Annotation Time and Quality

In Table 4 we summarized the annotation time and quality for three different corpora using the crosslingual clustering method. We also compared our approach with the human annotators who prepared for the ACE 2005 training corpora.

The annotation times in Table 4 are not directly comparable. The annotation time of our method covered the whole process of the five steps described in Section 4.2. For ACE, there were no available annotation time counts for individual event types; however, reportedly, it took about one year to finish defining and annotating all of the 33 event types. ACE training corpora produced more levels of annotations, which are not addressed in our task, such as event co-reference and event attributes. However, we can see that in general our method can provide a feasible way to discover and annotate a new event type. We also noticed that the annotation time spent on OntoNotes was considerably less than ACE corpora, mainly because the annotator used ground-truth SRL results and so saved a lot of time in argument correction.

The annotation task for the carbon sequestration literature was most efficient because the number of relevant sentences for the specified event type was the smallest among these three corpora, and also because the human annotator was a domain expert. Our initial concern for this domain was that the parsing and SRL models were both trained from the news domain and so may perform worse here, which in turn may incur additional annotation cost for correcting argument boundary and roles. However the results demonstrated that our approach can be extended to this domain in an effective and efficient way.

We also evaluated the annotation quality of the corrected corpora, against ground-truth (OntoNotes does not include event answer-keys so we only focused on the ACE and carbon sequestration corpora). It is encouraging to see that for the ACE corpus, our approach achieves comparable quality with the ACE human annotators. For example, in the following sentence:

Table 4. Event Annotation Time and Quality

Performance	Annotation Time	Trigger Labeling (%)			Argument Labeling (%)		
		P	R	F	P	R	F
ACE	12.7 hours	65.4	76.2	70.4	56.9	68.4	62.1
OntoNotes	6.2 hours	-	-	-	-	-	-
Carbon Sequestration	2 hours	100	100	100	82.7	82.7	82.7
ACE Human Annotator1	-	59.2	59.4	59.3	51.6	59.5	55.3
ACE Human Annotator2	-	69.2	75.0	72.0	54.1	73.7	62.4

Tehran had been governed by reformists since 1989, but ***[ARG1 a conservative city council]*** *was* ***elected*** *in the* ***[ARG-TMP February 28]*** *municipal polls in a result attributed to a meager turnout amid growing public disillusionment with electoral politics.*

All arguments involved in the “election” event were correctly labeled and so the annotator spent little time on this sentence and produced perfect event annotation after simple role mapping.

For the carbon sequestration corpus, the annotation quality was very high mainly because the ‘decrease’ events were not ambiguous in this domain and the argument identification was quite accurate. The main errors came from argument classification. For example, the human annotator needed to correct “*ARG1*” to “*ARG0*” for “*high flow rates*” in the following sentence:

Consequently while ***[ARG1 high flow rates] decrease [ARG1 the carbonation efficiency of the reservoir]*** *and low flow rates may reduce the permeability irreversibly close to the injection point moderate injection rates will ensure a partial carbonation of the rock and maintain the reservoir permeability.*

In these and many other cases, SRL failed to classify such arguments mainly because of some domain-specific features such as argument heads (e.g. *“rates”* appear very rarely as an “ARG0” in news domain). We expect to get further improvement after we incorporate some domain-specific knowledge such as high-frequency terminology lexicons into the SRL system.

6 Related Work

A number of previous studies have described extensive techniques to cluster words or word sequences from large unlabeled corpora (e.g. , Lin and Wu, 2009), monolingual parallel corpora (e.g. Lin and Pantel, 2001; Pang et al., 2003) and bilingual parallel corpora (e.g. Callison-Burch et al., 2008; Ji, 2009). Stevenson and Joanis (2003) applied semi-supervised learning for verb class discovery. We chose the parallel corpora discovery method and hybrid distributional clustering method because our target trigger list is a relatively closed set. Parallel corpora are likely to yield higher quality due to the human linguistic knowledge implicit in sentence alignment, but it is limited in size and vocabulary. Monolingual corpora are not as limited, so can cover more out-of-vocabulary terms, and might equal or out-perform parallel corpora methods if given a large enough monolingual corpus. Both methods do not require any supervision (beyond the sentence alignment for the bilingual method), so they are preferable for our purposes, both cost and time-wise.

Some event extraction systems demonstrated the positive impact of using SRL results (e.g. Grishman et al., 2005). There has been also some progress in addressing the portability issue of IE using completely automatic pipelines, such as on-demand IE (Sekine, 2006) and open-domain IE (Banko et al., 2007). We take a more modest approach because our ultimate goal is to generate a high-quality event annotation corpus for any newly discovered event type, and so that it can serve for the purpose of training an effective event tagger.

7 Conclusion and Future Work

In this paper we described some methods to address the portability issue of event extraction. We investigated this problem via two steps: (1). We automated the process of discovering novel event types by using two trigger clustering algorithms: (a) using a hybrid distributional semantic distance measure over a monolingual corpus, and (b) using statistical word-alignment over a bilingual parallel corpus; (2) We described a new view of event annotation based on semantic role labeling (SRL), and demonstrated that we can annotate event arguments efficiently and effectively by correcting SRL output. In order to evaluate the robustness of our approach, we experimented with both traditional news domain and a new domain of carbon sequestration. We expect that our method will enable developing annotated event corpora rapidly, and thus lead IE techniques to higher performance and broader applicability.

In the future we will focus on filtering the remaining noise in trigger clustering by adding more distributional constraints. In addition, we plan to use more advanced cost-conscious active learning methods (e.g. Haertel et al., 2008) to further speed up our annotation procedures, because our task fits naturally into a hierarchical framework with multiple annotation sub-tasks. We are also interested in applying domain adaptation techniques on SRL in order to boost event annotation quality for non-news domains.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Banko, Michele, Michael J Cafarella, Stephen Soderland and Oren Etzioni. 2007.Open Information Extraction from the Web. *Proc. IJCAI 2007*.

Callison-Burch, Chris. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. *Proc. EMNLP 2008*. Honolulu, USA.

Fellbaum, Christiane (Ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Fox, Christopher. 1992. Lexical Analysis and Stoplists. *Information Retrieval: Data Structures and Algorithms*. pp. 102-130.

Grishman, Ralph, David Westbrook and Adam Meyers. 2005. NYU's English ACE 2005 System Description. *Proc. ACE 2005 Evaluation Workshop*. Washington, US.

Grishman, Ralph. 2001. Adaptive Information Extraction and Sublanguage Analysis. *Proc. Workshop on Adaptive Text Extraction and Mining at Seventeenth International Joint Conference on Artificial Intelligence*.

Haertel, R., E. Ringger, K. Seppi, J. Carroll, and P.McClanahan. 2008. Assessing the costs of samplingmethods in active learning for annotation. *Proc. ACL* 2008.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. OntoNotes: The 90\% Solution. *Proc. HLT-NAACL 2006.*

Ji, Heng, Xiang Li, Angelo Lucia and Jianting Zhang. 2010. Annotating Event Chains for Carbon Sequestration Literature. Proc. LREC 2010.

Ji, Heng. 2009. Unsupervised Cross-lingual Predicate Cluster Acquisition to Improve Bi-lingual Event Extraction. *Proc. HLT-NAACL 2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics.*

LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. *http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf*

Lin, Dekang and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. *Proc. ACL 2009.*

Lin, Dekang and Patrick Pantel. 2001. DIRT-Discovery of Inference Rules from Text. *Proc. ACM SIGDD Conference on Knowledge Discovery and Data Mining.*

Marton, Yuval. 2010. Improved Statistical Machine Translation Using Monolingual Text and a Shallow Lexical Resource for Hybrid Phrasal Paraphrase Generation. *Proc. the Ninth Conference of the Association for Machine Translation in the Americas.*

Marton, Yuval, Chris Callison-Burch and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually derived Paraphrases. *Proc. EMNLP 2009*.

McDonald, Scott. 2000. *Environmental determinants of lexical processing effort*. Ph.D. thesis, University of Edinburgh.

Mohammad, Saif and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. *Proc. EMNLP 2006*.

Och, Franz Josef and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51.

Palmer, Martha, Daniel Gildea and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*. Volume 31, Issue 1. pp. 71-106.

Pang, Bo, Kevin Knight and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. *Proc. HLT/NAACL 2003*.

Pradhan, Sameer, Wayne Ward and James H. Martin. 2008. Towards Robust Semantic Role Labeling. *Computational Linguistics Special Issue on Semantic Role Labeling*, Vol. 34, No. 2, pp. 289-310, 2008.

Riloff, Ellen. 1996. Automatically Generating Extraction Patterns from Untagged Text. *Proc. AAAI 1996*

Sekine, Satoshi. 2006. On-Demand Information Extraction. *Proc. COLING-ACL 2006*.

Stevenson, Suzanne and Eric Joanis. 2003. Semi-supervised Verb Class Discovery Using Noisy Features. *Proc. CONLL 2003.*

Xue, Nianwen and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143-172.

Yangarber, Roman, Ralph Grishman, Pasi Tapanainen and Silja Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. *Proc. COLING 2000.*